Evaluation of the Vulnerability of Speaker Verification to Synthetic Speech

P.L. De Leon, M. Pucher, and J. Yamagishi. Evaluation of the vulnerability of speaker verification to synthetic speech. In Proc. Odyssey (The speaker and language recognition workshop) 2010, Brno, Czech Republic, 2010.

Evaluation of the Vulnerability of Speaker Verification to Synthetic Speech

Phillip L. De Leon *Michael Pucher* *Junichi Yamagishi*

New Mexico State University, Klipsch School of Elect. Eng., Las Cruces, New Mexico USA, pdeleon@nmsu.edu

Telecommunications Research Center (FTW), Vienna, Austria, pucher@ftw.at

Centre for Speech Technology Research (CSTR), Edinburgh, UK, jyamagis@inf.ed.ac.uk

Abstract

In this paper, we evaluate the vulnerability of a speaker verification (SV) system to synthetic speech. Although this problem was first examined over a decade ago, dramatic improvements in both SV and speech synthesis have renewed interest in this problem. We use a HMM-based speech synthesizer, which creates synthetic speech for a targeted speaker through adaptation of a background model and a GMM-UBM-based SV system. Using 283 speakers from the Wall-Street Journal (WSJ) corpus, our SV system has a 0.4% EER. When the system is tested with synthetic speech generated from speaker models derived from the WSJ journal corpus, 90% of the matched claims are accepted. This result suggests a possible vulnerability in SV systems to synthetic speech. In order to detect synthetic speech prior to recognition, we investigate the use of an automatic speech recognizer (ASR), dynamic-time-warping (DTW) distance of mel-frequency cepstral coefficients (MFCC), and previously-proposed average inter-frame difference of log-likelihood (IFDLL). Overall, while SV systems have impressive accuracy, even with the proposed detector, high-quality synthetic speech can lead to an unacceptably high acceptance rate of synthetic speakers.

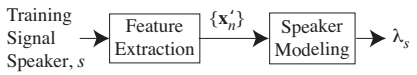
1. Introduction

The objective in speaker verification (SV) is to accept or reject a claim of identity based on a voice sample. During the training stage [Fig. 1(a)], speaker-dependent feature vectors $\mathbf{x}'_n$ are extracted from training speech signals and used to build a speaker model λ_s. The feature vectors are normally based on the mel-frequency cepstral coefficients (MFCCs). Within the speaker modeling block, feature vectors from all users are first concatenated and modeled with a Gaussian mixture model-universal background model (GMM-UBM) λ_{UBM} [1]. Next, the speaker model is constructed through MAP-adaptation of the GMM-UBM. Both λ_{UBM} and λ_s are parameterized by the set $\{w_i, \mu_i, \Sigma_i\}$ where w_i are the weights, μ_i are the mean vectors, and Σ_i are the diagonal covariance matrices of the GMM. During the testing stage [Fig. 1(b)], feature vectors $\mathbf{x}_n$ are extracted from a test signal and a log-likelihood ratio $\Lambda(\mathbf{X})$ is computed by scoring the sequence of test feature vectors $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ against the claimant model λ_C and the UBM

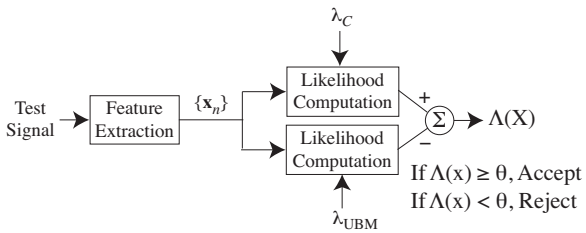
$$\Lambda(\mathbf{X}) = \log p(\mathbf{X}|\lambda_C) - \log p(\mathbf{X}|\lambda_{\text{UBM}}) \quad (1)$$

where

$$\log p(\mathbf{X}|\lambda) = \frac{1}{N}\sum_{n=1}^{N} \log p(\mathbf{x}_n|\lambda) \quad (2)$$



(a) Training stage.



(b) Testing stage.

Figure 1: GMM-UBM speaker verification system.

and N is the number of test feature vectors. The claimant speaker is accepted if

$$\Lambda(\mathbf{X}) \geq \theta \quad (3)$$

or else rejected, where θ is the decision threshold.

Synthetic speech potentially poses two related problems for SV systems. The first problem is confirmation of an acquired speech signal as having originated from a particular individual. In this case, the speech signal might be incorrectly confirmed as having originated from an individual when in fact the speech signal is synthetic. The second problem is in remote or on-line authentication where voice is used. In this case, a synthesized speech signal could be used to wrongly gain access to person's account. We assume for this second problem, the authentication system prompts the user to speak a randomly-chosen utterance in order to thwart the use of pre-recorded material, i.e. text-prompted SV. Of course, a randomly-chosen utterance would not present a problem for a speech synthesizer. In both of these problems, the speech model for the synthesizer must be targeted to a specific person's voice.

The problem of imposture against SV systems using speech synthesized from hidden Markov models (HMMs) was first published over 10 years ago by Masuko, et. al. [2]. In their original work, the authors used an HMM-based text-prompted SV system [3] and an HMM-based speech synthesizer. In the SV system, feature vectors were scored against speaker and background models composed of concatenated phoneme models (not GMM-based models). The authors also used a HMM-based speech synthesizer which was adapted to each of the human speakers [4].

When tested with 20 human speakers, the system had a 0% False Acceptance Rate (FAR) and 7.2% False Rejection Rate (FRR) and when tested with synthesized speech (20 synthetic voices) the system had over 70% FAR. In subsequent work by Masuko, et. al. [5], the authors extended the research in two ways. First, they improved their synthesizer by generating speech using pitch information. Second, they improved their SV system by utilizing both pitch and spectral information. The pitch modeling techniques used in synthesis were the same used in the SV system. By improving the SV system, the authors were able to lower the FAR for synthetic speech to 32%, however, the FAR for the human speech increased to 1.8%.

In the last 10 years, both speech synthesizers and SV systems have improved dramatically. Around the same time as Masuko's work, GMM-UBM-based SV systems were first proposed [6]. Since this time, GMM-UBM based SV systems have produced excellent performance and have achieved EERs of 0.1% on the TIMIT corpus (ideal recordings) and 12% on NIST 2002 Speaker Recognition Evaluations (SRE) (non-ideal recordings) [1, 7]. Other kernel-based techniques have been proposed and in some cases can lead to lower EERs, however, at this time GMM-UBM systems remain dominant in practice [8].

Until recently, developing a speech synthesizer for a targeted speaker required a large amount of speech data from a carefully prepared transcript in order to construct the speech model. However, with a state-of-the-art HMM-based speech synthesizer [9], the speech model can now be adapted from an average model (derived from other speakers) or a background model (derived from one speaker) using only a small amount of speech data. Moreover, recent experiments with HMM-based speech synthesis systems have also demonstrated that the speaker-adaptive HMM-based speech synthesis is robust to non-ideal speech data that are recorded under various conditions and with varying microphones, that are not perfectly clean, and/or that lack phonetic balance [10, 11]. In [11] a high-quality voice was built from audio collected off of the Internet. This data was not recorded in a studio, had a small amount of background noise, and the microphones varied in the data. Further [12, 13] reported construction of thousands of voices for HMM-based speech synthesis based on popular ASR corpora such as the Wall Street Journal (WSJ0, WSJ1, and WSJ-CAM0), Resource Management, Globalphone and SPEECON. Taken together, these state-of-the-art speech synthesizers pose major challenges to SV systems.

This paper is organized as follows. In Sections 2 and 3, we describe our speech synthesis and speaker verification systems. In Section 4, we describe the experimental evaluation and provide results using the Wall-Street Journal (WSJ) corpus and its synthesized counterpart. Although the WSJ journal corpus is not a standard corpus for SV research, it is one of the few that provides sufficient speech material from hundreds of speakers which is required to construct synthetic voices matched to their human counterparts. In Section 5, we investigate two methods for detecting synthetic speech and in Section 6, we examine the issue of duration adaptation in synthesized speech. We conclude the article in Section 7.

2. Speech Synthesizer

All text-to-speech (TTS) systems are built using the framework from the "HTS-2007 /2008" system [10, 14], which was a speaker-adaptive system entered for the Blizzard Challenge 2007 [15] and 2008 [16]. In the 2008 challenge, the system had the equal best naturalness and the equal best intelligibility on a training data set comprising one hour of speech. The system was also found to be as intelligible as human speech [14]. The speech synthesis system, outlined in Fig. 2, consists of four main components: speech analysis, average voice training, speaker adaptation, and speech generation.

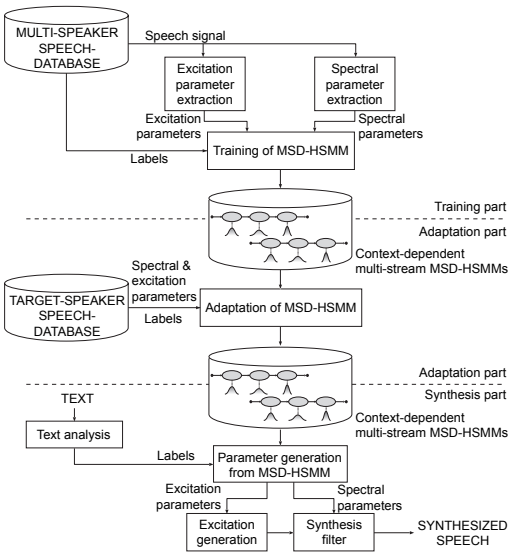


Figure 2: Overview of the HTS-2007/2008 speech synthesis system, which consists of four main components: speech analysis, average voice training, speaker adaptation, and speech generation.

In the speech analysis part, three kinds of parameters for the STRAIGHT (Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrogram [17]) mel-cepstral vocoder with mixed excitation (i.e., the mel-cepstrum, $\log F_0$ and a set of band-limited aperiodicity measures) are extracted as feature vectors for HMMs [18]. In the average voice training part, context-dependent multi-stream left-to-right multi-space distribution (MSD) hidden semi-Markov models (HSMMs) [19] are trained on multi-speaker databases in order to simultaneously model the acoustic features and duration. A set of model parameters (mean vectors and diagonal covariance matrices of Gaussian pdfs) for the speaker-independent MSD-HSMMs is estimated using the Expectation Maximization (EM) algorithm [20].

An overview of the training stages for the average voice models is shown in Fig. 3. First, speaker-independent monophone MSD-HSMMs are trained from an initial segmentation, converted into context-dependent MSD-HSMMs, and re-estimated. Then, decision-tree-based context clustering with the MDL criterion [21] is applied to the HSMMs and the model parameters of the HSMMs are tied at leaf nodes. The clustered HSMMs are re-estimated again. The clustering processes are repeated twice and the whole process is further repeated twice using segmentation labels refined with the trained models in a bootstrap manner. All re-estimation and re-segmentation processes utilize speaker-adaptive training (SAT) [22] based on constrained maximum likelihood linear regression (CMLLR) [23].

In the speaker adaptation part the speaker-independent

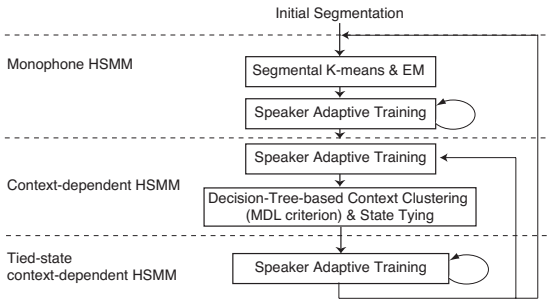


Figure 3: Overview of the training stages for average voice models.

MSD-HSMMs are transformed by using constrained structural maximum *a posteriori* linear regression (CSMAPLR) [24]. Note that not only output pdfs for the acoustic features but also duration models are also transformed in the speaker adaptation [25]. In the speech generation part acoustic feature parameters are generated from the adapted MSD-HSMMs using a parameter generation algorithm that considers both the global variance of a trajectory to be generated and trajectory likelihood [26]. Finally an excitation signal is generated using mixed excitation (pulse plus band-filtered noise components) and pitch-synchronous overlap and add (PSOLA) [27]. This signal is used to excite a mel-logarithmic spectrum approximation (MLSA) filter [28] corresponding to the STRAIGHT mel-cepstral coefficients to generate the speech waveform.

3. Speaker Verification System

The GMM-UBM SV system used in this research is shown in Fig. 1. Feature vectors are extracted every 10 ms using a 25 ms hamming window and composed of 15 MFCCs, 15 delta MFCCs, log energy, and delta-log energy as elements. We apply feature warping to the vectors in order to improve robustness [29] which is adequate given the high-quality recordings in the WSJ corpus. The GMM-UBM (1024 component densities) is built by concatenating the training feature vectors of the speakers within the corpus and using the EM algorithm to compute the parameters of the GMM-UBM. Individual speaker models are obtained through MAP-adaptation of the GMM-UBM (only the mean vectors) [1]. Our GMM-UBM SV system has a baseline for the 330 speaker NIST 2002 corpus (one speaker detection cellular task) of 12.10% EER which agrees with recently-published values [8]. In a previous study [30], our SV system was based on a Support Vector Machine (SVM) with Gaussian supervector kernel. Our results using a small, nonstandard corpus were the same as compared to the GMM-UBM system. In this paper, we have turned our attention to expanding the study with a much larger corpus than that used in [30] which implies the development of many more synthetic voices than previously used.

4. Experiments and Results

We use the WSJ corpus [31] from LDC. Although the WSJ corpus is not the de facto standard for building SV systems, it contains several hundred speakers and sufficiently long signals required for constructing each of the components of both our speech synthesizer and SV system [32]. Since the WSJ corpus mainly has relatively clean speech and there is no read for channel compensation or noise reduction, which would make

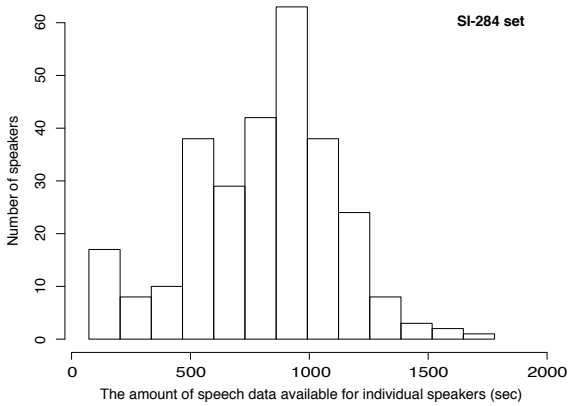


Figure 4: The amounts of speech data available for individual speakers in set A. The amounts vary from 73 sec to 27 mins.

distinction between real and synthetic speech more difficult due to the masking effects etc, we can expect results on ideal conditions. From the corpus, we chose the pre-defined official training data set (known as SI-284) that includes both WSJ0 and WSJ1 as material data. The SI-284 set has a total of 81 hours of speech data uttered by 284 speakers[1] was partitioned into three sets A, B, and C. Set A was used for the speech synthesis, i.e., constructing an average voice model and for adapting the average voice model to the 283 target speakers. Set B was used for constructing the SV system, i.e. constructing the UBM and adapting the UBM to the speakers. Note that the UBM and average voice model are trained on different subsets derived from the same corpus, since we aim to show results on the ideal conditions and thus should avoid cross-corpus negative effects. In future work, however, the UBM and average voice model will be derived from different corpora to discuss more practical situations where we also consider the cross-corpus effects including acoustic condition differences, microphone differences, or noise differences. Set C was used for test signals for the SV evaluation of human speech.

Since each speaker included in the SI-284 set has different speech durations, we used varying lengths (73 sec to 27 min) of training signals from set A to construct the average voice model and to adapt the model to each the speaker. Fig. 4 shows the distributions of the amounts of data available for individual speakers. The adapted models were used to create synthesized speech for each of the target speakers which serve as test signals for the SV evaluation of synthetic speech. Some speakers have larger amounts of data than those we can practically collect for the imposture against the SV system. However, in hopes that analyzing the quantity of data would give us some insight on this problem, we utilized the various amounts of data for speech synthesizer.

We used approximately 180 sec of material from each of the 283 speakers in set B for training and 30 sec of material from each of the 283 speakers in set C for testing. The Decision Error Tradeoff (DET) curve is shown in Fig. 5. With the decision threshold properly set, the EER is 0.4%. The mean and variance of the log-likelihood scores for the GMM-UBM SV system are computed and approximate score distributions for human speech are shown in Fig. 6 with green and red lines.

[1]Due to the recording condition issues 1 speaker was eliminated in our experiments.

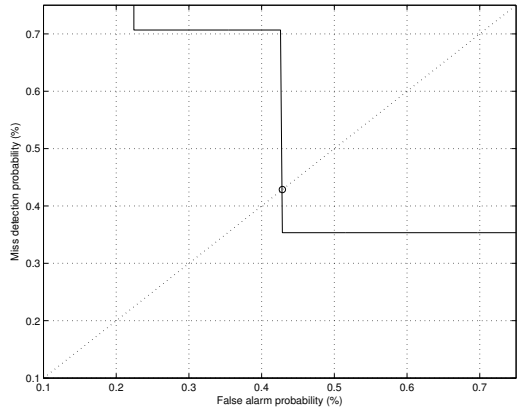


Figure 5: DET curve for speaker verification using test signals from human speakers. For the GMM-UBM system, the EER is 0.4%.

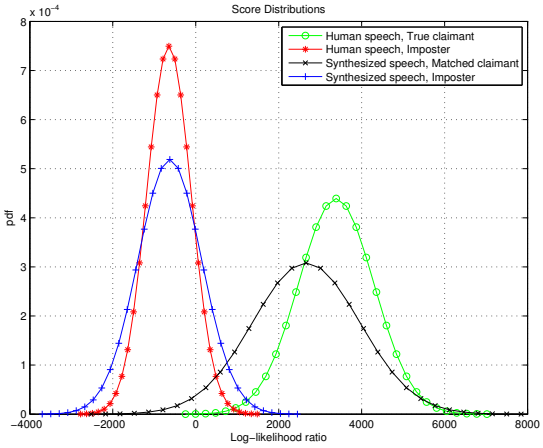


Figure 6: Approximate score distributions for GMM-UBM SV system with human and synthesized speech. Distributions for synthesized speech (black and blue lines) have significant overlap with those for human speech (green and red lines) leading to a 90% success rate for imposture using synthesized speech.

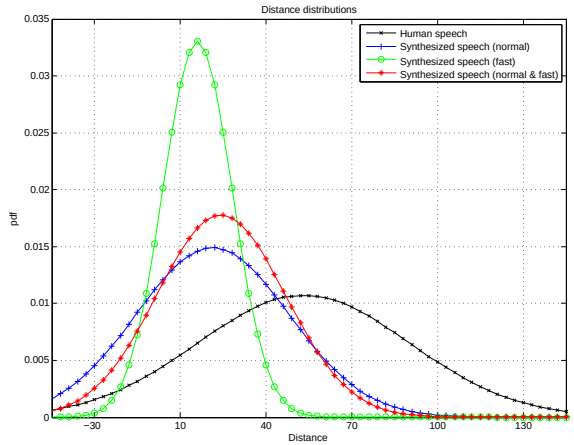


Figure 7: Distributions of DTW distance of MFCCs for human and synthetic speech with different linguistic contexts and durations for Austrian German corpus (9 speakers).

Using the same decision threshold, the system is then tested using synthetic speech. In this case, each of the 283 synthetic speech test signals is scored against one of 283 (human) claimant models leading to a total of 283^2 tests. Of these, 283 tests are with a "matched" claimant, i.e. synthesized voices claim to be their human counterparts and 283*282 tests with an "unmatched" or false claimant. For the matched claimant tests, 254/283 or 90% of the claims are accepted. Thus despite the excellent performance of the SV systems, the speaker similarity/identity of the synthesized speech is high enough to allow these synthesized voices to pass for true human claimants. The mean and variance of the log-likelihood scores (1) are computed and approximate score distributions for synthesized speech (black and blue lines) are shown in Fig. 6. As shown in Fig. 6 significant overlap occurs in the distributions of log-likelihood ratios for human speech, true claimant (green line) and synthesized speech, matched claimant (black line). Thus adjustments in decision thresholding or standard score normalization techniques are unlikely to differentiate between true and matched claims originating from human and synthesized speech [33, 34].

5. Detection of Synthesized Speech

The difference between a synthetic and human speech signal is audible although it is not clear at present, which acoustic cues are being used to discriminate. However, we have investigated three methods for the automatic (machine-based) detection of synthetic speech. The first uses a distance measure between MFCC features, the second uses the average inter-frame difference of log-likelihood (IFDLL) proposed in [35], and the third uses the word-error-rate (WER) and sentence-error-rate (SER) from an automatic speech recognizer (ASR) trained on human speech. With these three methods, we have designed two different system architectures for the detection of synthetic speech which are described in Subsection 5.4.

5.1. Dynamic time warping of MFCC features

In the first method, we compute the acoustic distance between two realizations of the same utterance using dynamic time warping (DTW) of MFCC features. This exploits the fact that the HMM-based synthesizer will always produce the same globally optimal waveform in terms of maximum likelihood, given a set of input phoneme labels while human speech will always be different. In Fig. 7, we see that synthetic speech phrases are more similar to each other (smaller variance) than human speech (larger variance), even when using different linguistic contexts and durations. This means that small changes in the synthesis parameters are not sufficient to make synthetic speech less similar [30].

5.2. Average inter-frame difference of log-likelihood

In the second method for detection of synthetic speech, we utilize the average IFDLL first proposed in [35]. The IFDLL is defined as

$$\Delta_n = |\log p(\mathbf{x}_n|\lambda_C) - \log p(\mathbf{x}_{n-1}|\lambda_C)| \quad (4)$$

and the average IFDLL is

$$\bar{\Delta} = \frac{1}{N}\sum_{n=1}^{N}\Delta_n. \quad (5)$$

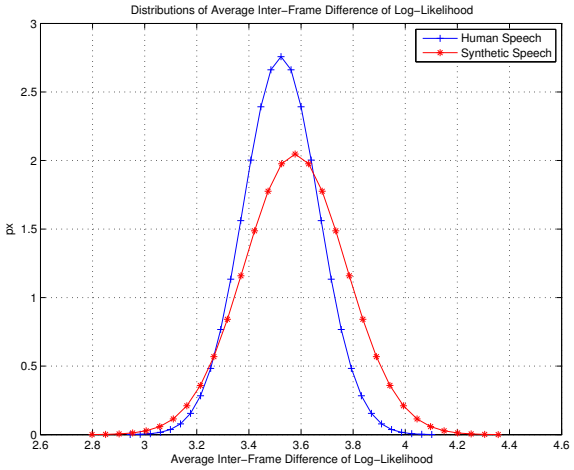


Figure 8: Distributions of average interframe-difference of log-likelihood for human and synthetic speech. Due to the overlapping distributions, the average IFDLL cannot be used to detect synthetic speech.

The authors in [35] observed that for synthetic speech, average IFDLL is lower than that for human speech. This difference was explained as a result of HMM-based synthesizers generating a speech parameter sequence so as to maximize the output probability. This maximization leads to a time variation of the speech parameters of synthetic speech becoming smaller than for human speech. In Fig. 8 we show the distributions of average IFDLL for human and synthetic speech using the 283 speaker WSJ corpus. Unlike the work in [35] which used the average IFDLL to detect synthetic speech, with state-of-the-art speech synthesis this measure no longer appears to be robust enough since the distributions in average IFDLL for human and synthetic speech have significant overlap. This can be explained because the state-of-the-art HMM-based speech synthesizer include global time variation models [26].

5.3. Automatic speech recognition

In the third method for detection of synthetic speech, we perform automatic speech recognition (ASR) on input utterances and examine WER. This can prevent some FAs from synthesizers trained with small amounts of speech as shown by the WERs and SERs in Table 1 taken from [30]. However, when we use all the training data available in the WSJ corpus (between 73 and 1620 sec per speaker) to train the synthesizer, WER for synthetic speech are *lower* than for real speech (Table 2). This means that we can use the method relying on WER only when the synthesizers, used by the impostor, are trained with very little data ($\approx$ 30 seconds). If enough training data is available, WER/SER does not appear a robust enough measure to detect synthetic speech. Grammars in Table 1 and 2 have a different number of possible input sentences that can be recognized using the grammar. This defines the grammars complexity, where the complexity of Grammar$[i]$ is lower than that of Grammar $[i+1]$.

In Fig. 9, we can see that the amount of training data used for the synthesizer does not impact the WER of the ASR. The amount of speech has to be significantly lower than that in Fig. 9 to significantly increase WER (see Table 1).

Table 1: Speech recognition WER/SER in % for Austrian German corpus (9 speakers) [30].

Dataset	Grammar1	Grammar2
Human speech	9.54 / 8.76	13.44 / 13.38
Synthetic (76 sec.)	11.64 / 10.82	15.62 / 16.09
Synthetic (38 sec.)	14.44 / 13.98	18.36 / 19.00
Synthetic (19 sec.)	26.50 / 29.52	31.33 / 36.16

Table 2: Speech recognition WERs and SERs in % for WSJ corpus (283 speakers).

Dataset	Grammar3	Grammar4
Human speech	9.55 / 10.79	13.91 / 33.24
Synthetic (73-1620 sec.)	3.05 / 4.85	5.50 / 22.51

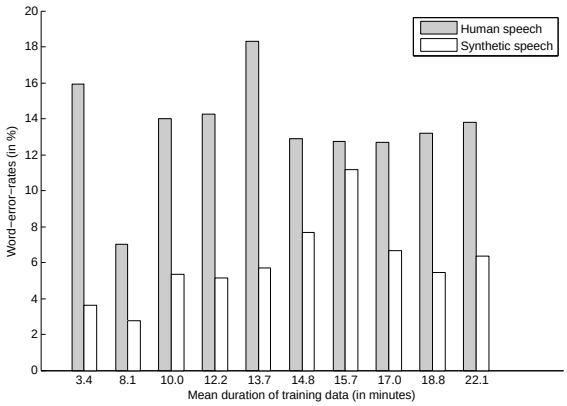
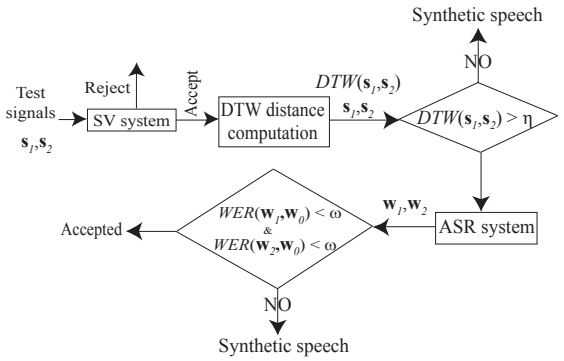
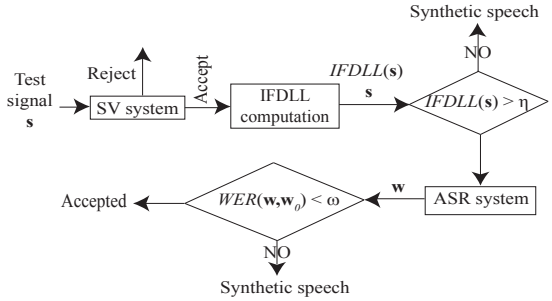


Figure 9: human and synthetic speech word-error-rates. Synthesizers are trained with different amounts of data from the Wall Street Journal corpus (283 speakers).



(a) System is composed of both MFCC DTW distance measures of repeated utterances and ASR.



(b) System is composed of both IFDLL measure and ASR.

Figure 10: Proposed systems for detection of synthesized speech after speaker verification.

5.4. Two systems for detecting synthetic speech

Although when used individually, the three methods for automatic detection of synthetic speech are not robust, they may be used in conjunction with one another to design different systems for detecting synthetic speech after SV acceptance. The first system shown in Fig. 10 (a) uses the DTW distance and ASR's WER. The input of the system are two signals, which are supposed to be realizations of the same utterances. The system relies on the high-degree of regularity of repeated utterances from a speech synthesizer and the higher error rates of an ASR (trained on human speech) subjected to synthesized speech trained on small amounts of speech data (see Table 1). If we have sufficient training data for the synthesizer we cannot use the WER for detecting synthetic speech (see Table 2). The system parameters are the distance threshold η, reference utterance word string $\mathbf{w}_0$, word-error-rate threshold ω, and distance function $DTW(\{\mathbf{x}_1\},\{\mathbf{x}_2\})$ based on mean or variance. After the ASR decoding it is verified that two utterances similar to the reference utterance were spoken. The similarity is defined by WER threshold.

The second system is shown in Fig. 10(b). It uses the IFDLL measure and ASR's WER. Here we only need one signal as input to the system. The system parameters are the IFDLL threshold η, reference utterance word string $\mathbf{w}_0$, word-error-rate threshold ω, and IFDLL function $IFDLL(\{\mathbf{x}_1\},\{\mathbf{x}_2\})$.

Our research has shown that the combination of these methods into the two different systems allows us to detect some examples of synthetic speech under certain circumstances (e.g. little training data available) as described in the previous sections. However accurate and reliable detection of synthetic speech is not yet achieved with these systems. For automatic detection of synthetic speech, it appears the general system designs of speech synthesizers must be taken into account. The measures and system architectures that we described here may serve as a starting point for building such a robust backend for the detection of synthetic speech.

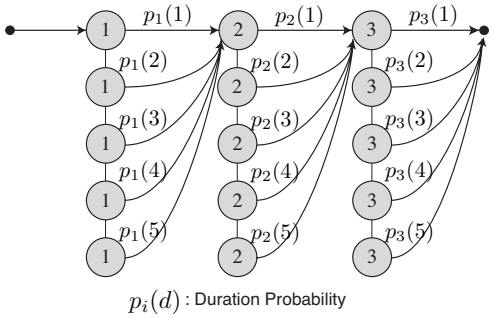


Figure 11: WFST-like illustration of duration models used for TTS systems. Duration probabilities p_i are also transformed to target speakers during speaker adaptation.

6. The Impact of Duration Adaptation in Synthesized Speech and Imposture

For speaker adaptation of speech synthesizer in the experiments so far, we transformed temporal structures of the HSMMs as well as mel-cepstrum and fundamental frequency streams to target speakers since the duration transformations perceptually improve synthetic speech [25]. However, the SV systems do not normally consider the temporal information and moreover they are known to be somewhat sensitive to recording condition mismatches and inconsistencies. In fact, the recording conditions of the WSJ corpus are not perfectly consistent and differ significantly among the recording sites [12].

Therefore in a second investigation we have turned off duration adaptation and compared it to synthesized speech with duration adaptation for evaluating the risk of speech synthesizers altered for the imposture purpose. The temporal/duration models used for our speech synthesizers are illustrated as Fig. 11, where we can see that each semi-Markov state has stack of states having associated duration probabilities p_i. The duration probabilities p_i are characterized by Gaussian pdfs, and the mean and variance of the pdfs are transformed to target speakers using CMLLR during speaker adaptation process. For the second investigation, we used the Gaussian pdfs without the CMLLR transforms for duration, which means that all speech synthesizers have same duration information derived from the average voice model, rather than speaker-specific duration information.

Fig. 12 and Table 3 show the results using IFDLL and ASR's WERs, respectively. As expected synthetic speech without duration adaptation has average IFDLLs closer to human and furthermore, lower WERs than those in Table 2 with duration adaptation. These additional results further convince us that with state-of-the-art speech synthesis, the previous and currently proposed methods to detect synthetic speech measure no longer appear to be robust enough.

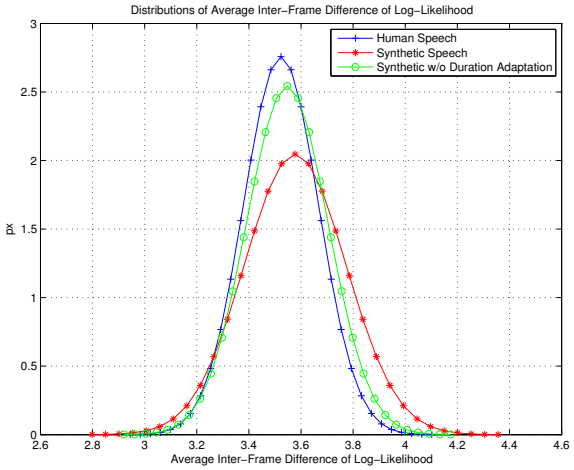


Figure 12: Distributions of average interframe-difference of log-likelihood for human, synthetic, and synthetic speech without duration adaptation. Due to the overlapping distributions, the average IFDLL cannot be used to detect synthetic speech.

Table 3: Speech recognition WERs and SERs in % for WSJ corpus without duration adaptation (283 speakers).

Dataset	Grammar3	Grammar4
Human speech	9.55 / 10.79	13.91 / 33.24
Synthetic (73-1620 sec.)	2.01 / 4.14	4.87 / 30.46

7. Conclusions

In this paper, we have evaluated the vulnerability of speaker verification (SV) to synthetic speech using state-of-the-art speech synthesis and SV systems using the relatively large Wall Street Journal corpus. Our results show that for matched claimant tests, 90% of the claims are accepted. Thus despite the excellent performance of the SV systems, the speaker similarity/identity of the synthesized speech is high enough to allow these synthesized voices to pass for true human claimants. This result suggests that high-quality synthetic speech may lead to a high false acceptance rate and may pose security issues for speech-based remote/online authentication or incorrect identity confirmation from a speech signal.

Next, we considered three measures for automatic detection of synthetic speech: 1) acoustic distance between two realizations of the same utterance using dynamic time warping (DTW) of MFCC features, 2) average inter-frame difference of log-likelihood, and 3) word-error-rate and sentence-error-rate from an automatic speech recognizer trained on human speech. Individually, these measures were found to not be robust enough to consistently (and with a high degree of accuracy) detect synthetic speech. We proposed two detection systems based on combinations of these measures.

Although automatic detection of synthetic speech appears to be a challenging problem, human listeners can easily perceive the difference when compared to human speech. Therefore our future work is to explore acoustic cues that we utilize in detecting these differences. Our goal is more robust 'code-breaking' features for the imposture of synthetic speech, based on perceptual analysis results and/or 'parallel' data of synthetic speech and real speech.

8. Acknowledgements

JY is partially supported by EPSRC and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project). This work has made use of the resources provided by the Edinburgh Compute and Data Facility which is partially supported by the eDIKT initiative (http://www.edikt.org.uk). The Telecommunications Research Center Vienna (FTW) is supported by the Austrian Government and the City of Vienna within the competence center program COMET.

9. References

[1] F. Bimbot, J F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleaua, S. Meignier, T. Merlin, J. Ortega-Garcia, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Applied Signal Process.*, vol. 4, pp. 430–451, 2004.

[2] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of HMM-based speaker verification systems against imposture using synthetic speech," in *Proc. EUROSPEECH*, 1999.

[3] T. Matsui and S. Furui, "Likelihood normalization for speaker verification using a phoneme- and speaker-independent model," *Speech Commun.*, vol. 17, no. 1-2, pp. 109–116, Aug. 1995.

[4] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," in *Proc. ICASSP*, 1996.

[5] T. Masuko, K. Tokuda, and T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch," in *Proc. ICSLP*, 2000.

[6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Dig. Sig. Process.*, vol. 10, pp. 19–41, 2000.

[7] T. Kinnunen, E. Karpov, and P. Franti, "Real-time speaker identification and verification," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, pp. 277–288, Jan. 2006.

[8] C. Longworth and M.L.F. Gales, "Combining derivative and parametric kernels for speaker verification," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 4, pp. 748–757, May 2009.

[9] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[10] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Trans. Speech, Audio & Language Process.*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[11] J. Yamagishi, Z.-H. Ling, and S. King, "Robustness of HMM-based speech synthesis," in *Proc. Interspeech 2008*, Brisbane, Australia, Sept. 2008, pp. 581–584.

[12] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, K. Oura, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for HMM-based speech synthesis," in *Proc. Interspeech 2009*, Brighton, UK, September 2009, pp. 420–423.

[13] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, K. Oura, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for HMM-based speech synthesis – Analysis and application of TTS systems built on various ASR corpora," *IEEE Trans. Speech, Audio & Language Process*, vol. in press, March 2010.

[14] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and K. Tokuda, "The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge," in *Proc. Blizzard Challenge 2008*, Sept. 2008.

[15] M. Fraser and S. King, "The Blizzard Challenge 2007," in *Proc. BLZ3-2007 (in Proc. SSW6)*, Aug. 2007.

[16] Vasilis Karaiskos, Simon King, Robert A. J. Clark, and Catherine Mayo, "The Blizzard challenge 2008," in *Proc. Blizzard Challenge Workshop*, Brisbane, Australia, September 2008.

[17] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[18] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.

[19] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.

[20] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[21] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Japan (E)*, vol. 21, pp. 79–86, Mar. 2000.

[22] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP-96*, Oct. 1996, pp. 1137–1140.

[23] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[24] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Speech, Audio & Language Process.*, vol. 17, no. 1, pp. 66–83, 1 2009.

[25] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[26] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[27] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–468, 1990.

[28] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP-92*, Mar. 1992, pp. 137–140.

[29] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. ODYSSEY*, 2001.

[30] P. L. De Leon, V. R. Apsingekar, M. Pucher, and J. Yamagishi, "Revisiting the security of speaker verification systems against imposture using synthetic speech," in *Proccedings of the 35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, USA, March 2010.

[31] Douglas B. Paul and Janet M. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*, Harriman, New York, 1992, pp. 357–362.

[32] "Wall Street Journal Corpus," 2010.

[33] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," *Proc. IEEE. Int. Conf. Acoustics, Speech and Signal Processing*, vol. 1, pp. 595–598, April 1988.

[34] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for test-independent speaker verification system," *Digital Signal Processing*, vol. 10, no. 1, pp. 42–54, 2000.

[35] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system," in *Proc. Eurospeech*, 2001.